Web-based Pseudo Relevance Feedback for Microblog Retrieval

Ahmed Saad El Din and Walid Magdy

Qatar Computing Research Institute
Qatar Foundation
Doha, Qatar
asaadaldien@hotmail.com, wmagdy@qf.org.qa

Abstract. This paper presents the experiments and results for the QCRI participation in the TREC Microblog track 2012. In this year, we apply a query expansion approach for improving the retrieval results in microblog search. Our approach performs web-search with the original query to get web results appeared at the same period of the query; then it extracts the webpage title of the first web result and uses it as an expansion to the original query before applying microblog search. Our results show a significant improvement to the baseline and significantly better results than the median result achieved in the track. We also report one run in the filtering task that applies straightforward technique for retrieval score threshold selection.

Keywords: Microblog search, Web-based query expansion, Twitter, TREC 2012

1 Introduction

This is the second year for running the Microblog track in TREC including the ad-hoc search task, which is concerned with searching tweets for relevant topics. In addition, a new filtering task was introduced this year. Our contribution was mainly directed to the ad-hoc search task, while we applied a straightforward approach for the filtering task.

Topical ad-hoc search is not the most popular search task on Twitter, since users mainly perform search to get updates about some entities or celebrities, find friends, get insight about certain hashtags … etc. [5]. Applying search for getting information on a given topic as in web search is not the most common task on Twitter because of the social nature of the domain, and the short length of the tweets that are focused and into the point. This makes searching for topics in Microblogs challenging but exciting at the same time.

Several approaches were introduced by participants of the track last year for performing an effective ad-hoc search on Twitter [4]. These approaches included: applying machine learning techniques for reranking results based on some tweets-specific features [2, 3], text normalization to Twitter language to convert slang

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2012		**00-00-2012 to 00-00-2012**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Web-based Pseudo Relevance Feedback for Microblog Retrieval	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Qatar Computing Research Institute,Qatar Foundation,Doha, Qatar,**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Presented at the Twenty-First Text REtrieval Conference (TREC 2012) held in Gaithersburg, Maryland, November 6-9, 2012. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA). U.S. Government or Federal Rights License

14. ABSTRACT
This paper presents the experiments and results for the QCRI participation in the TREC Microblog track 2012. In this year, we apply a query expansion approach for improving the retrieval results in microblog search. Our approach performs web-search with the original query to get web results appeared at the same period of the query; then it extracts the webpage title of the first web result and uses it as an expansion to the original query before applying microblog search. Our results show a significant improvement to the baseline and significantly better results than the median result achieved in the track. We also report one run in the filtering task that applies straightforward technique for retrieval score threshold selection.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**8**	

English words into proper language [6], applying query expansion using hashtags [1], and many other approaches [4]. Based on the results of last year, we noticed that the most effective approaches are those that seek enriching the queries/tweets with additional terms for better matching, and those that use large number of features for reranking the top retrieval results to increase the precision [2, 3]. We explored some of these techniques on the TREC 2011 data collection to investigate the effectiveness of each. Our preliminary results on the training set from 2011 showed that reranking approached did not achieve a significant improvement to the retrieval effectiveness, which does not align with the results achieved in the track last year. We did not further investigate this approach, since it requires much processing for extracting useful features, especially when the features requires additional crawling of information from Twitter and the web as those used in [3]. In contrast, we noticed from our experiments on the training set that pseudo relevance feedback (PRF) and query expansion in general always helps and usually leads to significantly better results. Therefore, we decided to focus our participation this year on query expansion techniques for the Microblog ad-hoc search task.

Regarding the newly introduced filtering task, we applied a straightforward approach for threshold selection based on maximizing the F-score.

The paper is organized as follows: Section 2 describes the collection of tweets and how it was preprocessed for indexing; Section 3 presents our main query expansion approach for the ad-hoc task; Section 4 shows our experimental setup and the submitted runs to the track; Section 5 reports the results; and finally, Section 6 concludes the paper.

2 Preparing Collection for Indexing

According to the track guidelines, only English tweets are considered relevant. Thus, we extracted the English from the approximately 16 million crawled tweets of the collection. We used the language-detection open source Java library[1]. A set of roughly 4.8 million English tweets were identified. We performed basic text tokenization where words were split on delimiters, except for "#" and "@" as they signify hashtags and user mentions respectively. All tweets starting with "RT", which indicates a retweet, were filtered out, since they are not considered relevant according to the track guidelines. This step reduced the number of English tweets to be indexed to almost 4.6 million tweets. Additionally, we removed all the URLs from the tweets text, since we assume that the URL text typically does not contain significant information in its own, and can cause some harm to the retrieval effectiveness.

The Indri search toolkit was used for indexing the collection of the 4.6 million tweets that were identified as English and non-retweets. Porter stemmer was applied during the indexing and search processes.

[1] http://code.google.com/p/language-detection/

3 Web-based Pseudo Relevance Feedback

The main challenge in finding relevant tweets to a given topic is the absence of sufficient word matching between the search query and the short tweet text. We applied some experiments on the TREC 2011 data and noticed the PRF always helps in improving the retrieval effectiveness of search, since it expands the query with additional terms that leads to better matching of more relevant tweets. In the approach we used this year, we attempt to apply query expansion to the query terms, but from external resources instead of the tweets collection itself.

Since the provided topics are all time-stamped, we believe that this time stamp of the topic is relevant to an event that occurred at that time, and the main objective is to find tweets discussing this event. Hence, our approach is to use the query provided for each topic to search the web at that time for relevant webpages (expecting to be news or articles discussing the topic), and use it to extract additional terms to enrich the query before searching the tweets collection.

A block diagram to our web-based query expansion technique is shown in Fig. 1. It works as follows:

1. The query of each of the topics is used to search Google at the time of the topic. In our experiments, the time was specified between the 25th of January 2011 (the earliest date of tweets in the collection) and the date of the topic itself. This assures that the returned results will be relevant to that topic at the time of issuing on Twitter.
2. The first web result in Google search is taken, which can be a news article, a forum, an online video … etc. The title of the result is then extracted.
3. The extracted title is then automatically processed by taking the first part of the title and dropping the second part that appears after any of the delimiters {"-", "|"}. These delimiters usually contain the web domain name of the webpage; e.g. "CNN.com", "YouTube", "Wikipedia".
4. The extracted titles of the pages are then appended to the original query for the search process in the tweets collection. Since some of the extracted titles are much longer than the original query, and to not dominate the weight of the original query, the weights of the original and the expanded parts of the query are set equal, and the final retrieval score is computed as the geometrical mean between the scores of the two parts.
5. An additional PRF step could be added to further apply query expansion from the collection of tweets itself.

Table 1 shows an example of the topics and the corresponding extracted webpages titles that were used for the expansion. As shown in Table 1, some of the extracted titles do not have any common words with original query. However, they are still relevant to the topic and leads to valuable enrichment to the query.

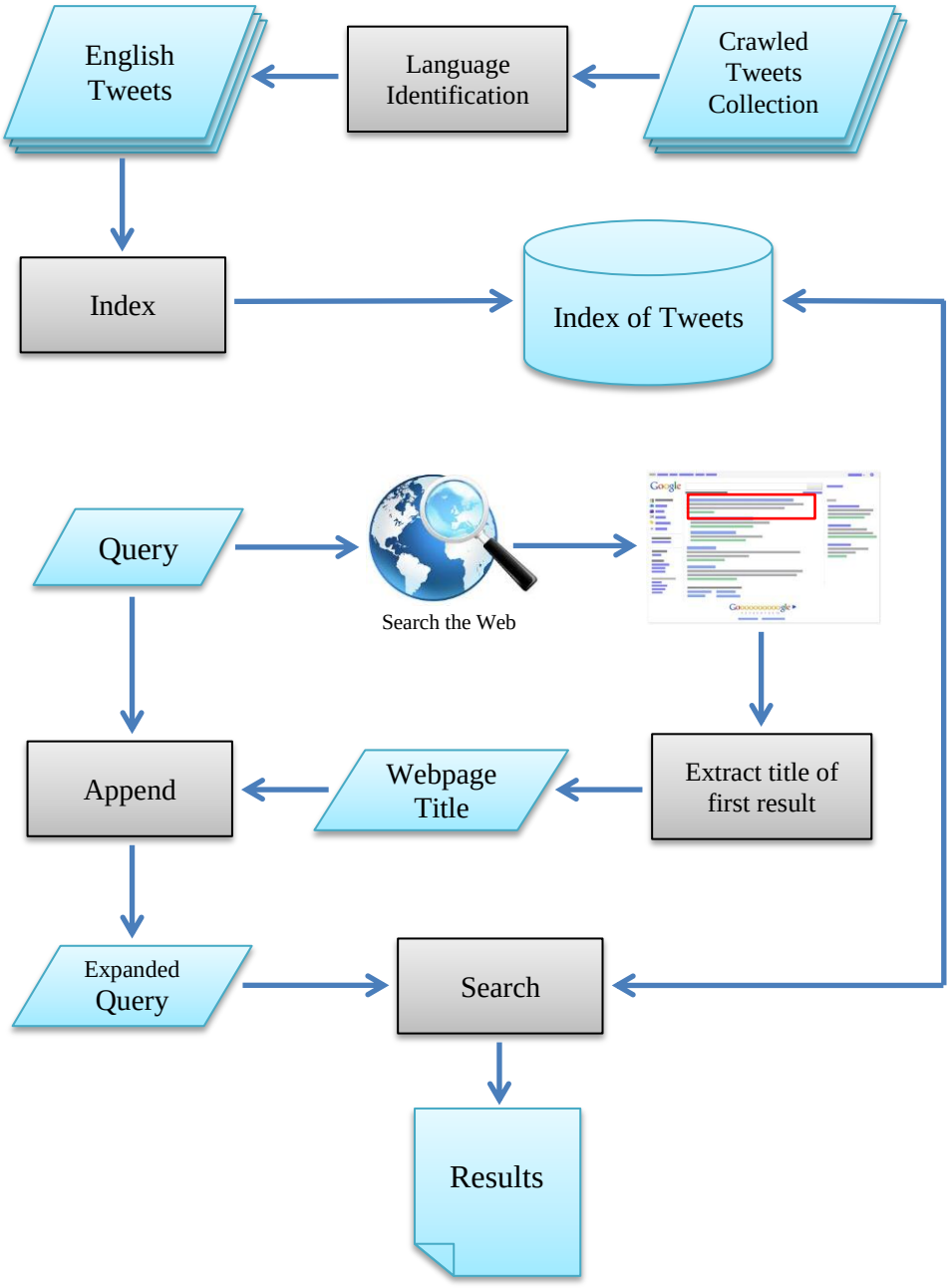

Fig. 1. Our approach for web-based query expansion in Microblog search

Table 1. Samples of the automatically extracted web-page titles of some of the topics

Topic ID	Query	Extracted Webpage Title
MB051	British Government cuts	UK Immigration UK Government Cuts Number of Skilled Occupations Eligible for Visa
MB056	Hugo Chavez	Why is Hugo Chavez called Dictator
MB063	Bieber and Stewart trading places	Justin Bieber Jon Stewart Switch Bodies On The Daily Show
MB076	Celebrity DUI violations	Drunk Driving and Hollywood
MB079	Saleh Yemen overthrow	Dictators Fall in the Middle East
MB092	stock market tutorial	Stock Market Tutorial Buying or Selling Stock with Strong Earnings
MB106	Steve Jobs' health	CEO Health Disclosure at Apple Public or Private Matter
MB109	Gasland	Mark Ruffalo Supported Documentary Gasland Nominated for Academy Award

4 Experimental Setup

Here we describe our submitted runs to the TREC 2012 microblog track. We submitted four runs to the ad-hoc task, and one run to the filtering task

4.1 Ad-Hoc Task

The four runs submitted to the track are as follows:

1. **BL**: The baseline run, where the queries of topics were taken without any modifications and used to search the collection. Indri was used for search.
2. **BLFB**: The same as baseline, but PRF was applied. The number of documents used for the feedback process was 50 tweets, and the number of expansion terms was set to 10.
3. **QEWeb**: Query expansion using the web was applied as discussed in pervious section.
4. **QEWebFB**: Similar to QEWeb, but PRF was applied.

4.2 Filtering Task

The submitted run for the filtering task applied a straightforward algorithm for detecting a threshold of the retrieval score to be used for filtering. The retrieval algorithm we used for this task used the same setup of the ad-hoc run QFWEBFB, since it showed the best results on our training data.

For detecting the retrieval score threshold (S_0) for the filtering process, we modelled the probability of relevant document given the retrieval score $P(r|s)$ and the probability of the non-relevant document given the retrieval score $P(n|s)$ as two normal distributions. We then inferred the parameters values using 80% of the training topics (8 topics in our case) of the training set and we use the remaining topics (2 topics) for tuning other optimization parameters. Initially, we maximized the precision by maximizing the expected number of the relevant document by setting it to its minimum value, where $P(r|s) = P(n|s)$. Later, we optimized the threshold by selecting the one that leads to the highest F-score.

Algorithm:

1- Set $\alpha = 1$.
2- Find S_0 $where:$ $\int_{S_0}^{1} P(r|s)ds > \alpha \int_{0}^{S_0} P(n|s)ds$
3- $\alpha = \alpha + 0.01$
4- If F-Score increased go to 2 else return S_0

4.3 Evaluation

For the evaluation, we report the P@30, MAR, R-Prec, and ROC curves for our four submitted runs. These score are computed when considering only the highly relevant tweets (relevance>2). In addition, we computed the scores when considering all the relevant tweets (relevance > 0).

For the filtering run, scores are reported as received from the track, which include four scores: precision, recall, F_0.5, and t11su.

5 Results

Table 2. P@30, MAP, and R-Prec for the four submitted runs when considering only the highly relevant tweets (rel. = 2) vs. when considering all relevant tweets (rel. > 0) for the ad-hoc search task

	Highly Relevant			**All Relevant**		
	P@30	**MAP**	**R-Prec**	**P@30**	**MAP**	**R-Prec**
BL	0.1701	0.1512	0.1838	0.3141	0.2191	0.2666
BLFB	0.1718	0.1638	0.1884	0.3169	0.2326	0.2784
WEB	0.1881	0.1706	0.1988	0.3548	0.2531	0.2910
WEBFB	**0.1921**	**0.1710**	**0.2001**	**0.365**	**0.2548**	**0.3074**

Table 3. Results of the filtering task

Precision	**Recall**	**F_0.5**	**t11su**
0.3571	0.4651	0.3436	0.3245

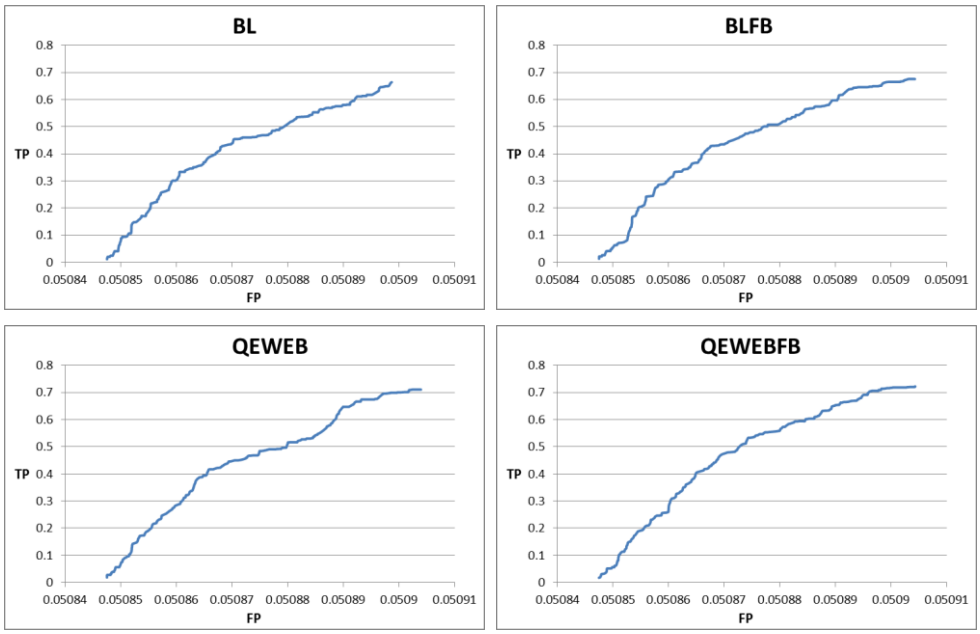

Fig. 2. ROC curves for our four runs for the ad-hoc search task

Table 2 shows the retrieval results of our runs for the ad-hoc task. It is very clear how query expansion leads to improvement to the retrieval effectiveness. The additional information added to the query from search the web led to a significant improvement compared to the baseline. Also, PRF leads to further improvement to the retrieval effectiveness. Figure 2 presents the ROC curves of the four runs.

Table 3 reports our result for the filtering task. We are not sure how our technique performs compared to other algorithms, but we know that our scores is significantly higher than median achieved scores in the track.

6 Conclusion and Future Work

We can conclude our paper by suggesting when doing topical ad-hoc search for microblogs to apply query expansion from the web, which proved to enrich the query with additional valuable information that improves the search results significantly.

For future work, we aim to further investigate this technique of expansion by exploring different mechanisms for extracting the expansion terms rather than just extracting the webpage title.

7 References

1. A. El-Kahki, K. Darwish. QCRI @ TREC 2011 : Microblog Track. In TREC-2011
2. D. Metzler, C. Cai. (2011). USC/ISI at TREC 2011: Microblog Track. In TREC-2011.
3. Z. Obukhovskaya, K. Pervyshev, A. Styskin, P. Serdyukov. Yandex at TREC 2011 Microblog Track. In TREC-2011
4. I. Ounis, C. Macdonald, J. Lin, I. Soboroff. (2011). Overview of the TREC-2011 Microblog Track. In TREC-2011.
5. J. Teevan, D. Ramage, M. Morris. #Twittersearch: A comparison of microblog search and web search. *WSDM 2011*.
6. Z. Wei, L. Zhou, B. Li, K.-F. Wong, W. Gao, K.-F. Wong. (2011). Exploring Tweets Normalization and Query Time Sensitivity for Twitter Search. In TREC-2011